Install Pyspark

In [1]:
```
!pip install pyspark
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/publ
ic/simple/
Collecting pyspark
  Downloading pyspark-3.2.1.tar.gz (281.4 MB)
     |████████████████████████████████| 281.4 MB 33 kB/s
Collecting py4j==0.10.9.3
  Downloading py4j-0.10.9.3-py2.py3-none-any.whl (198 kB)
     |████████████████████████████████| 198 kB 41.7 MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.2.1-py2.py3-none-any.whl size=281853642
sha256=f2686fb889fbe1813d68af554ef1a8228ef3daf58a57f3c972b749190245ab42
  Stored in directory: /root/.cache/pip/wheels/9f/f5/07/7cd8017084dce4e93e84e92efd1e1d53
34db05f2e83bcef74f
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.3 pyspark-3.2.1
```

Create Session

The entry point to programming Spark with the Dataset

In [2]:
```
from pyspark.sql import SparkSession
```

In [3]:
```
session = SparkSession.builder.appName("sql").master("local").getOrCreate()
```

In [5]:
```
data = session.read.csv("churn.csv", header = True, inferSchema = True)
```

In [7]:
```
data.show(5)
```

```
+----------+------+-------------+-------+----------+------+-----------+-----------------
--+------------------+-------------------+-------------------+-----------------------+--
-----------------+-------------------+-------------------+--------------+-------------+-
---------------+---------------------+-----------+-----+
|customerID|gender|SeniorCitizen|Partner|Dependents|tenure|CallService|MultipleConnectio
ns|InternetConnection|     OnlineSecurity|       OnlineBackup|DeviceProtectionService|
    TechnicalHelp|           OnlineTV|       OnlineMovies|     Agreement|BillingMethod|
  PaymentMethod|MonthlyServiceCharges|TotalAmount|Churn|
+----------+------+-------------+-------+----------+------+-----------+-----------------
--+------------------+-------------------+-------------------+-----------------------+--
-----------------+-------------------+-------------------+--------------+-------------+-
---------------+---------------------+-----------+-----+
|2907-ILJBN|Female|          0.0|    Yes|       Yes|  11.0|        Yes|
No|                No|No internet service|No internet service|    No internet service|No
internet service|No internet service|No internet service|      One year|           No|
  Mailed check|                 20.6|      233.9|   No|
|3896-RCYYE|Female|          0.0|     No|        No|  67.0|         No|   No phone servi
ce|               DSL|                 No|                 No|                    Yes|
              Yes|                Yes|                Yes|Month-to-month|          Yes|
   Credit card |                 53.4|    3579.15|   No|
|9764-REAFF|Female|          0.0|    Yes|        No|  59.0|        Yes|
No|                No|No internet service|No internet service|    No internet service|No
internet service|No internet service|No internet service|      Two year|           No|
Bank transfer |                 18.4|    1057.85|   No|
|6651-RLGGM|  Male|          0.0|    Yes|       Yes|  67.0|        Yes|                Y
es|                No|No internet service|No internet service|    No internet service|No
internet service|No internet service|No internet service|      Two year|           No|
  Mailed check|                    26.3|      1688.9|    No|
|5879-SESNB|Female|            0.0|      No|          No|   11.0|         Yes|                 Y
es|       Fiber optic|                   No|                    No|                        No|
               No|                  No|                    No|Month-to-month|              No|E
lectronic check|                   75.25|       888.65|    No|
+-----------+------+-------------+-------+----------+------+-----------+------------------
--+-------------------+-------------------+-------------------+-----------------------+--
-----------------+-------------------+-------------------+--------------+-------------+-
---------------+---------------------+-----------+-----+
only showing top 5 rows
```

Q1 What is the average Monthly Charges for customers having “DSL” Internet connection. Consider senior citizen , male and churned customers whose tenure is greater than 60.

In [28]:

```python
First = data.filter((data['InternetConnection'] == 'DSL') & ((data['gender']== 'Male') &
```

```
+--------------------------+
|avg(MonthlyServiceCharges)|
+--------------------------+
|         76.05276750166666|
+--------------------------+
```

Q2 What is the average amount for customers having “Mailed Check” as payment method and “One Year” as agreement. Consider the customers who have dependents and partner and have opted for call service.

In [11]:

```python
from pyspark.sql.functions import mean
second = data.filter((data['PaymentMethod'] == 'Mailed check') & (data['Agreement']=='On
```

```
+------------------+
|  avg(TotalAmount)|
+------------------+
|1208.3856143694081|
+------------------+
```

Q3 What is the total Monthly Service Charges of customers having different billing method. Consider the male and senior citizen customers whose tenure is less than 20 and have multiple connections

In [31]:

```python
from pyspark.sql.functions import sum
third = data.filter((data['BillingMethod']=='Yes') & (data['gender']== 'Male') & (data['
```

```
+--------------------------+
|sum(MonthlyServiceCharges)|
+--------------------------+
|             7906.29044558|
+--------------------------+
```

Q4 How many male and female customers has dependents and no dependents. Consider those customers who have monthly service charges greater than 100

In [22]:

```python
from pyspark.sql.functions import count ,when,col
fourth = data.filter(data['MonthlyServiceCharges'] >100).groupBy("gender").agg(count(whe
```

```
+------+--------------------------------------------------+-------------------------------
-------------------+
|gender|count(CASE WHEN (Dependents = Yes) THEN True END)|count(CASE WHEN (Dependents =
No) THEN True END)|
+------+--------------------------------------------------+-------------------------------
-------------------+
|Female|                                               183|
                438|
|  Male|                                               182|
               389|
+------+-------------------------------------------------+-------------------------------
------------------+
```

Q5 How many number of customers have churned and not churned. Consider only female customers who have no dependents and has done call service and has preferred electronic check method.

In [33]:
```python
from pyspark.sql.functions import count
fifth = data.filter((data['gender']=='Female') & (data['Dependents']=='No') & (data['Cal
```

```
+--------------------------------------------+-------------------------------------------
-+
|count(CASE WHEN (Churn = Yes) THEN True END)|count(CASE WHEN (Churn = No) THEN True EN
D)|
+--------------------------------------------+-------------------------------------------
-+
|                                         718|                                         56
6|
+--------------------------------------------+-------------------------------------------
-+
```

Q6 How many male and female customers have no dependents and have multiple connections. Consider the customers who have call service and has preferred either electronic check method or mailed check method

In [24]:
```python
from pyspark.sql.functions import count
sixth = data.filter((data['CallService'] == 'Yes') & ((data['PaymentMethod'] == 'Electro
```

```
+------+------------------------------------------------+--------------------------------
--------------------------+
|gender|count(CASE WHEN (Dependents = No) THEN True END)|count(CASE WHEN (MultipleConnec
tions = Yes) THEN True END)|
+------+------------------------------------------------+--------------------------------
--------------------------+
|Female|                                            2068|                                
                      1252|
|  Male|                                            2071|                                
                      1233|
+------+------------------------------------------------+--------------------------------
--------------------------+
```

Q8 What is the maximum monthly service charges of customers who have done payment by electronic check method? Consider only those customers who have agreement for on year or two years only.

In [25]:
```python
from pyspark.sql.functions import max
eight = data.filter((data['PaymentMethod'] == 'Electronic check') & ((data['Agreement'] 
```

```
+--------------------------+
|max(MonthlyServiceCharges)|
+--------------------------+
|                    118.65|
+--------------------------+
```

Q9 What is the minimum total amount of male and female customers having one year or two year agreement type. Consider only those customers who have no internet connection, no online security no online backup and no device protection service.

In [27]:
```python
from pyspark.sql.functions import min
ninth = data.filter(((data['Agreement'] == 'One year') | (data['Agreement'] == 'Two year
```

```
+------+----------------+
|gender|min(TotalAmount)|
+------+----------------+
|Female|     69.21978888|
|  Male|     59.02463086|
+------+----------------+
```

In [ ]:

```
```